FEMS Microbiology Reviews, fuw017

doi: 10.1093/femsre/fuw017
Review Article

REVIEW ARTICLE

Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes

Miklós Bálint[1,*], Mohammad Bahram[2,3], A. Murat Eren[4,5], Karoline Faust[6], Jed A. Fuhrman[7], Björn Lindahl[8], Robert B. O'Hara[1], Maarja Öpik[2], Mitchell L. Sogin[4], Martin Unterseher[9] and Leho Tedersoo[10]

[1]Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25, 60325 Frankfurt, Germany, [2]Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai St., 51005 Tartu, Estonia, [3]Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden, [4]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA, [5]Department of Medicine, University of Chicago, Chicago, IL 60637, USA, [6]Center for the Biology of Disease, Rega Institute, KU Leuven, 3000, Belgium, [7]Department of Biological Sciences, University of Southern California, MC0371, Los Angeles, CA 90089, USA, [8]Department of Soil and Environment, Swedish University of Agricultural Sciences, Box 7014, SE-750 07, Uppsala, Sweden, [9]Institute of Botany and Landscape Ecology, Ernst-Moritz-Arndt University Greifswald, Soldmannstr. 15, Greifswald, D-17487, Germany and [10]Natural History Museum, University of Tartu, 14a Ravila St., 50411 Tartu, Estonia

*Corresponding author: Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25, D-60325 Frankfurt, Germany. Tel: +49-(0)69-7542-1856; E-mail: mbalint@senckenberg.de

One sentence summary: This is an overview of the more widely adopted and emerging techniques for analysis of diversity and community composition, and the inference of species interactions from co-occurrence data generated by high-throughput sequencing of marker genes.

Editor: Jan Roelof van der Meer

ABSTRACT

With high-throughput sequencing (HTS), we are able to explore the hidden world of microscopic organisms to an unprecedented level. The fast development of molecular technology and statistical methods means that microbial ecologists must keep their toolkits updated. Here, we review and evaluate some of the more widely adopted and emerging techniques for analysis of diversity and community composition, and the inference of species interactions from co-occurrence data generated by HTS of marker genes. We emphasize the importance of observational biases and statistical properties of the data and methods. The aim of the review is to critically discuss the advantages and disadvantages of established and emerging statistical methods, and to contribute to the integration of HTS-based marker gene data into community ecology.

Keywords: marker gene; community ecology; microbial ecology; microbial diversity; species interactions; statistical methods

Received: 30 December 2015; **Accepted:** 23 May 2016

INTRODUCTION

Communities of microorganisms—bacteria, archaea, fungi and protists—are highly diverse and drive globally important ecosystem functions (van der Heijden *et al.* 1998; Leininger *et al.* 2006; Falkowski, Fenchel and Delong 2008; van der Heijden, Bardgett and van Straalen 2008). Beyond their ecological importance, microbial communities are also notoriously difficult to comprehensively describe. For a long time, cultivation-based approaches were the only available methods to characterize microbial communities, leading to the discovery of novel phyla and classes from diverse habitats (Stingl *et al.* 2008; Margesin and Miteva 2011; Rosling *et al.* 2011). However, as the majority of microorganisms are uncultivable by standard techniques (Epstein 2013), molecular methods such as Sanger sequencing of cloned products and PCR amplicons, PCR-RFLP, were adopted for their identification and classification. High-throughput sequencing (HTS) techniques were successfully used in microbial community ecology soon after their development (Sogin *et al.* 2006). HTS initiated a new era of microbial ecology, characterized by a considerable increase in data volume and an urgent need for bioinformatics skills. HTS methods allow us to test new ecological hypotheses that require hundreds of samples and/or high taxon coverage.

HTS-based biodiversity analyses have inherent problems that are continuously being revisited to improve the accuracy of the biological inferences (Quince, Curtis and Sloan 2008; Kunin *et al.* 2010). Critical issues include sequencing errors, chimeric sequence formation during PCR (Wang and Wang 1997) and sequencing library preparation (Carlsen *et al.* 2012), as well as preferential amplification of taxa due to primer bias (Sipos *et al.* 2007; Tedersoo *et al.* 2015). These problems are not exclusive to HTS, but accumulate due to the sheer amounts of data typical of HTS studies. Several bioinformatics tools and work flows address quality filtering, sequence clustering and identification (Schloss *et al.* 2009; Abarenkov *et al.* 2010; Caporaso *et al.* 2010; Bálint *et al.* 2014) and provide sequence count matrices of operational taxonomic units (OTUs) that serve as inputs for statistical analyses.

Buttigieg and Ramette (2014) emphasized three key reasons why microbial ecologists should have a broad overview of statistical methodology: (i) statistical methods are frequently re-evaluated and updated to match analytical needs; (ii) ongoing debates constantly add new perspectives on how to approach certain types of data; and (iii) the rapid evolution of both molecular and statistical methods provide continuous challenges and opportunities for analysis. Since many of the analytical tools cannot cope with huge data matrices (Jansson and Prosser 2013), our review focuses on HTS-capable approaches. Our purpose is to provide a critical overview and guidance for the design and statistical analysis of HTS-based community surveys. We discuss pros and cons of both established and emerging statistical tools. We focus on new approaches that have yet to penetrate microbial ecology, particularly multispecies model-based statistical methods. Such models are more flexible and easier to interpret than commonly used methods, and have better statistical properties for large community matrices (Hui *et al.* 2015; Warton *et al.* 2015b). Nonetheless, we refrain from establishing 'gold standards', because these may hamper scientific creativity and further methodological improvements. Finally, we illustrate model-based approaches to the analysis of community composition with data from a global soil fungal data set, and species interaction inference from time-lagged co-occurrences with a long monthly marine time series (Gilbert *et al.* 2012; Tedersoo *et al.* 2014). The analysis code is provided through a GitHub repository, a resource that can be dynamically kept up-to-date by the community, while still preserving the possibility to curate content. The repository is accessible via https://github.com/MikiBalint/micro-ecol-tools.git.

SAMPLING DESIGN

Establishing an appropriate sampling design is the first critical step in community ecology, because it greatly affects the statistical power of analyses and the reliability of the results and their interpretation. Above all, sampling design must be appropriate for the hypotheses being tested. The availability of resources, ecosystem type and the biology of target organisms need to be considered when optimizing the number of sampling units, their spatial and temporal arrangement, and the size of individual samples. These aspects will influence the representativeness, the spatial and temporal independence of observations. Many books on biometrics and community ecology summarize the principles of study design for hypothesis testing and provide ample examples of good practices with appropriate data structure and replication levels for both field surveys and manipulative experiments (e.g. Quinn and Keough 2002; Legendre and Legendre 2012; Manly and Alberto 2014). Although originally designed for the analysis of communities of macroorganisms, most of these approaches are also applicable to the analysis of microbial community data (Prosser *et al.* 2007).

GENERAL PROPERTIES OF HTS COMMUNITY DATA

What are the units of observation of HTS community data, and how are these generated? The way that the data are generated will affect the analysis and results, and several aspects specific to HTS community data need to be considered.

Generation of sequence counts

Sequence counts represent the primary unit of observation in HTS community ecology. These data are generally used in a similar way as individual abundances of different species in community ecology of macroorganisms. Ideally, sequence counts should reflect marker gene counts in cells, but the biases involved in read generation considerably distort the counts. For example:

(i) The number of marker gene copies per cell varies among organisms, with the exact numbers remaining unknown in most cases (Eickbush and Eickbush 2007; Bik *et al.* 2013).
(ii) The efficiency of DNA extraction depends on the structure of cell walls. PCR may introduce quantitative biases due to primer-template mismatches and length difference of amplicons (Ihrmark *et al.* 2012; Parada, Needham and Fuhrman 2015; Tedersoo *et al.* 2015a).
(iii) PCR may generate artificial base changes (Brodin *et al.* 2013) and chimeric molecules (Ashelford *et al.* 2006). The use of proofreading DNA polymerase enzymes may reduce the incidence of PCR errors several-fold (Oliver *et al.* 2015), but simultaneously increases the incidence of chimeras (Schnell, Bohmann and Gilbert 2015).
(iv) Further, chimera formation known as tag switching may occur between different samples, if amplicons from different samples are combined in the library preparation step (Carlsen *et al.* 2012; Schnell, Bohmann and Gilbert 2015).

Table 1. Major resources for bioinformatics and taxonomic identification of microorganisms, and dedicated bioinformatics software.

SILVA	Ribosomal RNA sequences	http://www.arb-silva.de/	Pruesse *et al.* (2007)
Ribosomal database project	Ribosome-related data and services, including RNA sequences, derived phylogenetic trees, etc.	http://rdp.cme.msu.edu/	Cole *et al.* (2009)
GreenGenes	16S gene database	http://greengenes.lbl.gov/	DeSantis *et al.* (2006)
UNITE	Fungal ITS sequences	http://unite.ut.ee/	Kõljalg *et al.* (2013)
MaarjAM	18S, ITS and 28S sequences for the arbuscular mycorrhizal fungi	http://maarjam.botany.ut.ee/	Öpik *et al.* (2014)
Mothur	Bioinformatics tool for marker gene data	www.mothur.org	Schloss *et al.* (2009)
QIIME	Bioinformatics pipeline for marker gene data	http://qiime.org/	Caporaso *et al.* (2010)
VAMPS	Visualization and analysis tools for microbial communities	https://vamps.mbl.edu/	Huse *et al.* (2014)
Megan	Taxonomic analysis of BLAST results	http://ab.inf.uni-tuebingen.de/software/megan5/	Huson *et al.* (2011)
SCATA	Bioinformatic pipeline for non-alignable sequences, e.g. ITS	http://scata.mykopat.slu.se	

The impact of these biases differs among studies, and the investigators should judge how much particular biases may influence the conclusions drawn from the data. For example, results relying on relative abundance may be distorted due to primer bias, richness estimates are influenced by PCR and sequencing errors, and tag switching can result in false positive observations. It seems impossible to evaluate the relative importance of the biases as the combination of underlying factors is unique to every study system. Probably the best approximation can be obtained with taxonomically complex positive controls ('mock communities', see below), and numerous negative controls (extraction, PCR, tagging negatives), included into all phases of sequence generation.

Definition of ecologically meaningful units

Marker gene amplicon sequences are commonly grouped into OTUs to provide a culture-independent method to characterize microbial community structures in environmental samples. These OTUs are generated either with reference-based or *de novo* approaches, or a combination of both (Bik and Thomas 2012). The goal of both reference-based and *de novo* approaches, regardless of their use of sequence similarity cut-offs, is to identify sequences in datasets that can act as proxies for a population of microbial genomes in a sample. Several marker gene databases are commonly used for protists, bacteria, archaea and fungi. Multiple programs are available that can combine the enormous information content of these databases with an array of simple analytical techniques (Table 1). Connecting sequences to an established taxonomic framework may be useful, because the community can then be described with reference to existing knowledge, e.g. in terms of the functional guilds present (Fierer, Bradford and Jackson 2007; Clemmensen *et al.* 2015). This is often of higher priority for eukaryotes, where studies commonly aim to identify taxonomically recognized species, and where marker gene analyses developed closer to taxonomy due to the more common occurrence of recognizable macrostructures and the relatively easy use of a biological species concept. Translation between OTUs and biological species is not straightforward, and the term 'species hypothesis' may be used to stress that grouping of sequences aims to reflect biological species, yet with some uncertainty (Kõljalg *et al.* 2013). In bacteria and archaea, the lack of a unified species concept (Gevers *et al.* 2005; Doolittle and Zhaxybayeva 2009) and difficulties associated with cultivation and archiving type material (Burbank and Anderson 2012) result in their astonishing underrepresentation in taxonomic databases (Pace 1997; Sogin *et al.* 2006; Hinchliff *et al.* 2015). Consequently, the community analysis of bacteria and archaea is more detached from finer level taxonomy, which limits the number of different kinds of microbial organisms that can be taxonomically identified in HTS data sets.

Taxonomy- and reference-independent (*de novo*) approaches delimit OTUs by comparing each read in a data set to other reads in the same data set. These are purely bioinformatic procedures and do not necessarily involve considerations of the nature of a (biological) species or taxonomy. The most common strategy for *de novo* OTU delimitation uses algorithms that form clusters with a sequence similarity cut off (which is often set at 97%–99% for bacteria, archaea and fungi; Huse *et al.* 2010). OTUs identified through *de novo* approaches can often be linked to a taxonomic framework through their similarity to sequences of known taxonomic origin; thus, they can inherit taxonomic assignments based on their closest similarity to entries in a reference database of previously characterized organisms. A special type of *de novo* OTU delimitation uses evolutionary models to infer OTUs either with a phylogenetic species concept (reviewed by Fujita *et al.* 2012), or with Poisson tree processes (Zhang *et al.* 2013).

Although clustering based on sequence similarity is widely used, there is a growing appreciation that the 'standard' 97% sequence similarity threshold often fails to identify ecologically relevant and/or phylogenetically unmixed units of bacteria, archaea (Koeppel and Wu 2013; Patin *et al.* 2013; Tikhonov, Leach and Wingreen 2015) and eukaryotes (Ryberg 2015). For example, oligotyping analysis of bacterial rRNA amplicon sequences from coastal waters of Cape Cod, Massachusetts showed that two *Pelagibacter* oligotypes with 99.6% nucleotide identity at the

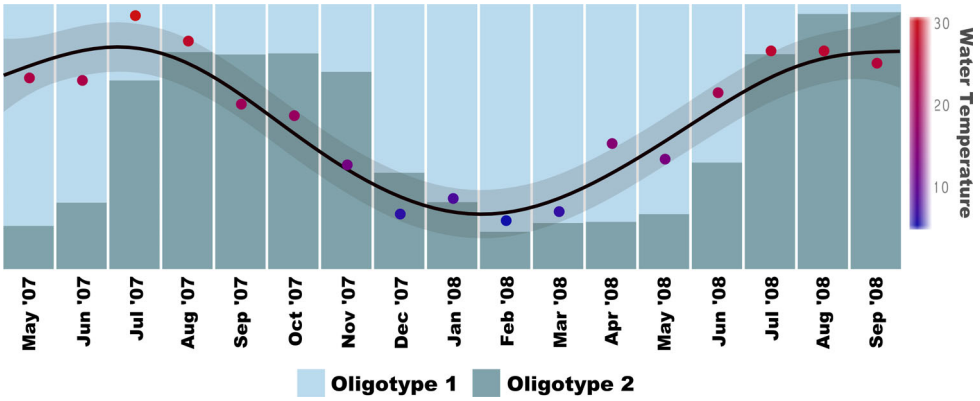

Figure 1. Seasonal variation of two *Pelagibacter* oligotypes that are 99.6% identical at the V4–V6 region of the 16S rRNA gene based on their relative abundance (Eren *et al.* 2013).

sequenced region, fluctuated remarkably with seasonal changes in water temperature (Fig. 1; Eren *et al.* 2013). Thus, several bioinformatics algorithms have been developed that do not require a global sequence similarity cut-off.

(i) Clustering 16S rRNA for OTU Prediction (CROP; Hao, Jiang and Chen 2011) is an unsupervised Bayesian clustering method operating on the natural organization of data, without hard cutoff thresholds.
(ii) Distribution-Based Clustering (DBC; Preheim *et al.* 2013) identifies ecologically distinct groups by combining the distribution patterns of sequences across samples with their genetic distances. DBC relies on genetic distance in conjunction with co-occurrence patterns of studied sequences to identify OTUs.
(iii) The agglomerative clustering algorithm SWARM (Mahé *et al.* 2014) uses pairwise sequence similarities to form single-linkage clusters of sequences that occur close to each other in sequence-distance space.
(iv) Minimum Entropy Decomposition (MED; Eren *et al.* 2015a) iteratively decomposes a given data set using only highly variable nucleotide positions to create highly resolved units.

Different OTU delimitation approaches may result in different OTU sets that may affect the results and interpretation (Schmidt, Matias Rodrigues and von Mering 2015). Phylogenetic information may also have considerable effects on OTU delimitation, emphasizing the limitations of sequence similarity-based approaches (Nguyen *et al.* 2016b). *De novo* OTU delimitation outperformed reference-based approaches in a recent comparison (Westcott and Schloss 2015), but only reference-based approaches produce stable OTUs (He *et al.* 2015). More comparative studies, which include reference-based and *de novo* approaches (with or without hard sequence similarity thresholds, and phylogenetic considerations), are needed. Existing comparisons are regularly done on real-life data from diverse sequencing projects. There is much need for comparisons done on simulated data, which allows for complete control over the sequences and a better evaluation of the algorithm performances.

Analytical consequences of data generation and processing

Although the relative abundance of organisms is of great importance, quantitative interpretations are complicated by analytical biases. If samples are processed similarly, comparisons of OTU abundance among samples may be informative, because any methodological biases should be the same (Amend, Seifert and Bruns 2010). The absolute abundance of microbes may be estimated by scaling read numbers to positive extraction controls, created by adding a controlled amount of DNA from several taxa during DNA extractions (Smets *et al.* 2015). Although positive controls will not help to identify the reasons why a species has more reads compared to an other, they may help to evaluate the extent of biases in read abundances among taxa. The improving knowledge about copy number variation across taxa through genomic studies (Perisin *et al.* 2016) combined with quantification of DNA could provide a better picture about the actual abundance of taxa. *In silico* approaches (Ficetola *et al.* 2010) and mock communities (Ihrmark *et al.* 2012; Parada, Needham and Fuhrman 2015) can be useful to evaluate the correspondence of true and observed abundances. However, the results of *in silico* primer evaluations should be treated carefully, as these results may only moderately correlate with sequenced mock community composition (Parada, Needham and Fuhrman 2015). Competition between templates during PCR differs by samples, and this adds an unknown amount of error to every PCR run, even if a standardization aliquot is used. The abundance of selected species may be evaluated with real-time PCR, but results will be also influenced by multiple confounding factors, such as the AT/GC ratio, marker gene fragment length, marker gene copy numbers, etc. Real-time PCRs on single copy genes and metagenomic approaches are probably the least biased ways to evaluate abundances.

HTS typically generates many OTUs represented by a single or few reads. While many of these rare OTUs are undoubtedly genuine entities (Eren *et al.* 2015b), a large fraction of rare sequence types are generated through PCR, sequencing or sequence processing errors (Quince *et al.* 2009; Parada, Needham and Fuhrman 2015). Artificial taxa are particularly problematic for richness estimation and community analysis using binary (presence/absence) data, because rare OTUs are similarly weighted as more abundant ones. Therefore, many researchers remove global singletons (i.e. OTUs that were found only once in the entire dataset) and consider OTUs with a few total counts as unreliable. For ultra-HTS Illumina data sets, OTUs with less than 5–10 sequences are frequently removed (Brown *et al.* 2015). More conservatively, such pruning may be conducted on a per-sample basis, i.e. OTUs with local abundance under a specific threshold are removed from the analysis. It is arguable whether OTUs should be removed from samples, where they are represented

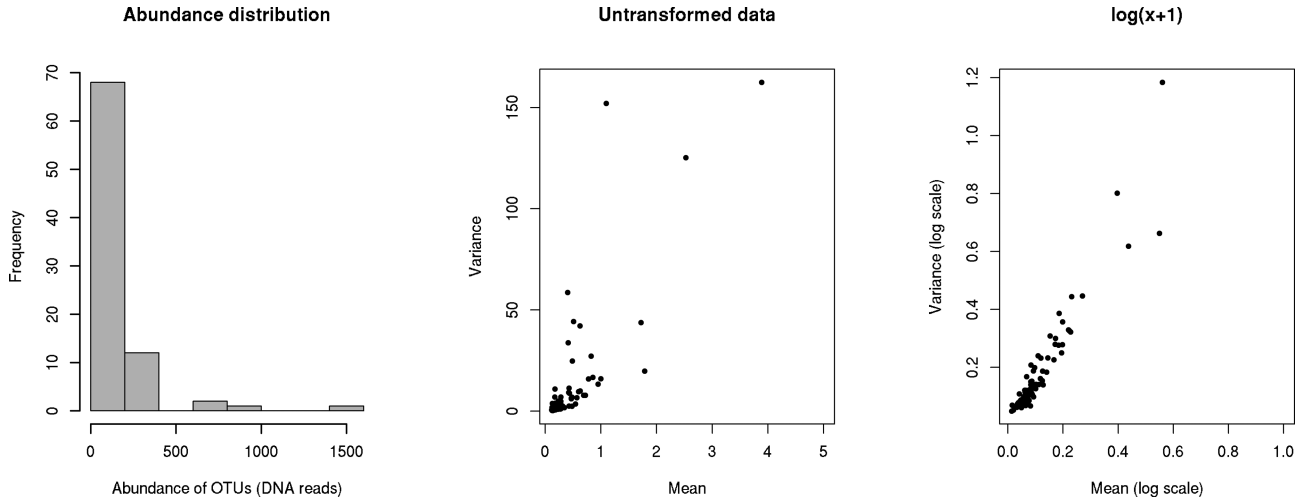

Figure 2. Statistical properties of HTS marker gene data. (**A**) Distribution of plant pathogen OTU abundances. Most OTUs were represented by very few sequence numbers, with a few OTUs being represented by large sequence counts. (**B**) Heteroscedasticity in the untransformed data. (**C**) Heteroscedasticity is still present in the log-normalized counts. Data represents pathogenic soil fungal OTUs, extracted from a global data set on soil fungi (Tedersoo *et al.* 2014).

by only a few reads in one sample, while they are abundant in other samples. Pruning may also be connected to the sequencing depth by applying a relative abundance threshold to define rare OTUs.

HTS read numbers may vary by orders of magnitude among individual samples. This is generally problematic in downstream analyses, because the sequencing efforts are almost always positively correlated with observed richness, especially if there are large differences in sequencing effort among the samples (Haegeman *et al.* 2013). It is important to consider the effects of differential sampling in all analyses, as many analytical methods assume the same sampling effort for all samples. In the worst case, patterns in composition and diversity may simply reflect the differential sequencing depth among the samples. Sequence numbers are frequently normalized by rarefying: random subsampling to a common sequencing depth. Rarefaction is analytically problematic and poses multiple statistical problems: (i) omission of available valid data, (ii) the estimation of overdispersion is more difficult due to data loss, (iii) loss of power (type II error), (iv) dependence on an arbitrary threshold and (v) additional uncertainty due to the randomness in rarefaction (McMurdie and Holmes 2014). Rarefaction may eventually lead to a loss of pattern in rare, but biologically relevant OTUs, because OTUs with low read numbers are more likely to be removed from samples. Such frequent, but low-abundance OTUs may be extremely interesting ecologically (Ainsworth *et al.* 2015; Eren *et al.* 2015b). Quantification of the relative importance of these rarefaction-related problems warrants further research. Another approach is to convert the sequence counts into relative abundances by dividing OTU counts with the total number of reads observed in each sample. Richness, nonetheless, remains higher in samples that were more deeply sequenced. Although there are more positive views on rarefaction (e.g. Weiss *et al.* 2015), several approaches avoid rarefaction and normalization altogether, e.g. by directly incorporating sequencing depth differences into all analyses (Bálint *et al.* 2015). The relative importance of the sequencing depth versus the variables of interest can then be quantified and used as an indicator of the reliability of the results (Bálint *et al.* 2015). Other alternatives rely on variance-stabilizing transformations developed for RNA-Seq data, such as TMM (Robinson and Oshlack 2010) and RLE (Risso *et al.* 2014), implemented in the R packages edgeR (Robinson, McCarthy and Smyth 2010) and DESeq2 (Love, Huber and Anders 2014). Weiss *et al.* (2015) provide a comparison of available data normalization techniques, except the direct incorporation of sequencing depth into the analysis.

Because HTS marker gene counts are discrete, a Gaussian statistical distribution is usually not an appropriate model (Warton, Wright and Wang 2012; Haegeman *et al.* 2013; Bálint *et al.* 2015). Taxonomic abundance matrices are characterized by many zeroes and a few large counts (Fig. 2A). Further, the mean-to-variance ratio is not stable, i.e. the variation of observation frequencies of a sequence type is dependent on the frequency of observation (heteroscedasticity; Fig. 2B). Exponential growth rates of organisms are one biological reason for heteroscedasticity: the log(mean) versus log(variance) increases with a slope of 2 when sampling from an exponential distribution, and with a slope of 1 when sampling from a Poisson distribution (Taylor's Law; Taylor 1961). Not considering heteroscedasticity may lead to (i) confounding location and dispersion effects, (ii) low detection power of a multivariate effect expressed in low-variance taxa and iii) difficulties in detecting the taxa that drive particular effects (Warton, Wright and Wang 2012). OTU abundances are often transformed using a square root, fourth root or logarithmic function to address the typical non-normality of ecological counts, but these transformations often fail to stabilize variances (Fig. 2C).

ESTIMATION OF RICHNESS, AND COMPARISON OF RICHNESS AND DIVERSITY OF SAMPLES

Many—if not most—species remain unobserved during sampling due to the aggregated distribution of organisms, insufficient sampling effort and a wide range of sampling biases (Chao *et al.* 2009; Colwell *et al.* 2012). Thus, the total species richness is often estimated in the course of biodiversity studies (i) as an estimate of the true richness and completeness of sampling, and (ii) to evaluate richness differences among samples with differential sampling or sequencing depth. It is widely acknowledged that richness estimators are unreliable at small sample sizes (Chao *et al.* 2009). Statistically, reliable estimations are only possible if either the true number of individuals in the community

is known or the distribution of abundances is known: otherwise it is unclear what proportion of the individuals are used in the richness estimation (O'Hara 2005). If the number of individuals in a community is high, sampling is always limited to a small fraction of the individuals in the community, leaving most of the rare species not sampled.

Species richness estimators depend on the rare species to estimate the number of even rarer species. For HTS data, this is problematic, because an unknown proportion of rare sequences only observed once or twice are artifacts. The typical practice of discarding rare OTUs (which are defined either with sequence counts, or by incidence rarity—presence in a few samples only; e.g. Bálint *et al.* 2015) is particularly incompatible with richness estimators as they rely on the abundance distribution of the rarest species.

A large number of non-parametric and parametric richness estimators have been developed for macroorganism count- and sample-based data (e.g. Chao1, Chao2, Jacknife1, Jacknife2, ICE and ACE) and introduced to microbial ecology (Hughes *et al.* 2001; Bohannan and Hughes 2003; Unterseher *et al.* 2008). Non-parametric estimators account for only the parts of the community that were actually sampled (representative spatiotemporally) and therefore provide only a lower bound for the true richness. Parametric estimators assume that the observed abundances follow some distribution (e.g. lognormal or gamma species abundance distribution) and estimate the number of unsampled species accordingly (Hong *et al.* 2006; Locey and Lennon 2016). Taxon sampling curves are generated either by adding or removing samples with different sizes independently and sequentially (Gotelli and Colwell 2001). Species richness is estimated by fitting a curve to these samples, and predicting the richness at some asymptote. These estimates are implicitly parametric, because the shape of the curve is assumed *a priori*, and is a function of the observed abundance distribution (O'Hara 2005). Simulations, where the number of individuals and species abundance distributions are controlled show that true richness can be estimated only with parametric estimators and if the species abundance distributions are correctly defined (O'Hara 2005). Given the problems presented in the previous paragraph, richness estimation should be avoided in HTS studies.

Richness, or some other measure of diversity, is often compared among samples to evaluate the effect of the environment on communities, although species these measures are not nearly as important as community composition in detecting the effects of environmental change (Magurran *et al.* 2015; Magurran 2016). Comparison of sample richness requires comparable observation efforts among samples (i.e. similar sequencing depth). The general solution is to compare samples either with extrapolated (see problems above), or interpolated richness. Recently, Chao and Jost (2012) proposed an integrated approach of interpolation and extrapolation to compare species richness. During interpolation, species accumulation or rarefaction curves and their confidence intervals are used to assess differences in richness (Colwell, Mao and Chang 2004). However, differences inferred through interpolations are only reliable if species abundance distributions of compared communities are the same (McGill *et al.* 2007; Gwinn *et al.* 2016).

The problem of the unknown number of individuals is less important if other aspects of diversity are used, which discount low-sequence-count OTUs. Such comparisons can be done with diversity indices (e.g. Shannon, Simpson) that weight rare and common taxa differently. Rényi diversity series and the corresponding Hill numbers (Hill 1973) are visualized in curves, where the weight of rare species varies along a continuous scale, and include Shannon and Simpson's diversities as special cases. However, diversity indices are not straightforward to interpret, because any observed difference depends on the definition of diversity. This needs to be discussed when reporting such comparisons instead of simply stating that one sample is more diverse than another. A detailed overview about measuring biodiversity was written by Magurran (2003).

ANALYSIS OF COMMUNITY COMPOSITION

Research on microbial community composition targets four questions as follows. (i) What are the similarities in OTU occurrence? (ii) Is there a change in community composition because of an experimental treatment, environmental variation, etc.? (iii) What are the causes (predictors) of change? (iv) Which OTUs are most affected in the community? The main complication for answering these questions lies in the multivariate nature of both species observations, and the numerous predictors that influence the communities. Potential collinearities in both OTU observations and among predictors further complicate these analyses.

Similarities in OTU occurrence

Similarities in OTU occurrence and/or sample properties are traditionally identified by a wide range of ordination tools (Quinn and Keough 2002; Legendre and Legendre 2012). Common examples include non-metric multidimensional scaling (NMDS), principal component analysis (PCA), principal coordinate analysis (PCoA) and correspondence analysis (CA). Most ordination tools arrange objects in multidimensional space either after being supplied with, or after computing a distance matrix. Distance matrices can be calculated based on multiple alternative distances or dissimilarity metrics, but typical biological count data exhibit overdispersion, i.e. the variance is larger than would be expected, assuming typical distributions such as a Poisson (Warton, Wright and Wang 2012; Anderson and Walsh 2013). As a consequence, locality and dispersion effects may not be distinguished (see below). The Euclidean distance of Hellinger-transformed data (Hellinger distance) and Bray–Curtis distance are most widely used in plant ecology and microbial ecology, including HTS data sets. A general discussion of distance and similarity measures is available from Legendre and Legendre (2012), and a discussion specific to marker gene-based microbial community studies is provided by Kuczynski *et al.* (2010). If communities differ strongly in taxonomic richness (e.g. along the environmental gradients), most distance metrics are affected by richness differences or the nested component of diversity rather than species replacement (Chase *et al.* 2011). Although nestedness may be used as an indicator of environmental effects, it is generally useful to treat these components separately. The betapart method (Baselga 2010; Baselga and Orme 2012) partitions beta diversity into nested and turnover components. An other solution is to use indices that account only for shared species, such as the modified Raup–Crick metric or beta.SIM value (Chase *et al.* 2011). The drawback of these metrics is reliance on binary data and thus sensitivity to treatment of rare species and false positives due to sequencing artifacts. As presence–absence measures overestimate the importance of rare species in calculating similarities, it is important to use the weighted version of these indices because abundant OTUs may have greater influence on community functions (such as the probabilistic version of the Bray-Curtis dissimilarity; Stegen *et al.* 2013).

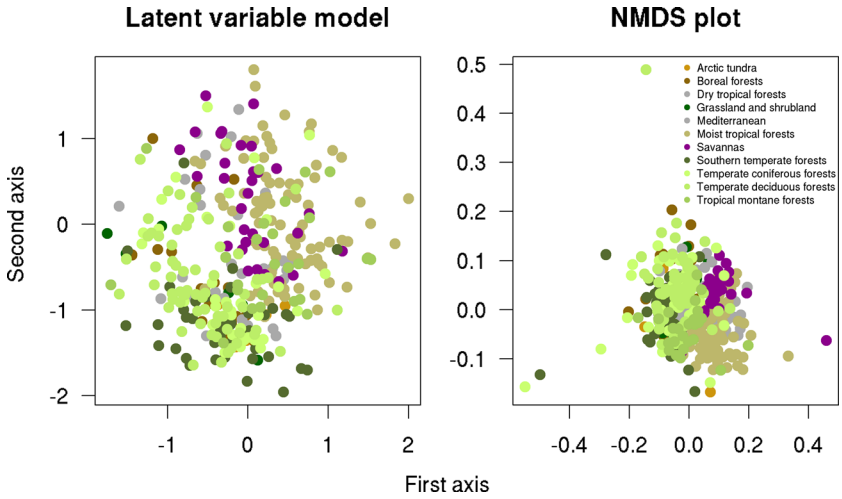

Figure 3. NMDS and model-based ordination of plant pathogen sequence counts across different biomes (Tedersoo *et al.* 2014). The latent variable model accounts for overdispersion and retains location effects.

Phylogenetic distance metrics address questions related to the evolutionary history of a community, or phylogenetic conservatism versus phylogenetic overdispersion (i.e. community members are phylogenetically more closely or distantly related than expected by chance). These metrics combine phylogenies with abundance data to address evolutionary changes in communities. Common metrics include the mean pairwise distance (MPD) and mean nearest neighbor distance (MNND), that are based on the average and nearest neighbor distances across a phylogenetic tree (Vamosi *et al.* 2009). Pairwise phylogenetic distances reveal phylogenetic relationships at deep nodes, while nearest neighbor distances such as MNTD are sensitive to the extent of clustering towards tips. The UniFrac distance family relies on average neighbor distances and it is designed to cope with uncertainties in species assignments or OTU delimitations (Lozupone and Knight 2005; Lozupone *et al.* 2007). A simplified, less resource-intensive version is also available (Hamady, Lozupone and Knight 2009). The phylogenetic context in which these distances are computed may be considered with edge principal component analysis and squash clustering (Matsen and Evans 2013). Highly variable and thus difficult to align markers such as ITS and trnF require alternative phylogenetic hypothesis construction, e.g. cluster analyses, and manual mapping onto a phylogenetic backbone.

Correspondence between distance matrices can be evaluated with Mantel tests and Procrustes analyses. The statistical properties of the Mantel test have been criticized for overestimating the degrees of freedom, and thus elevating type I error rates (Guillot and Rousset 2013). The Procrustes analysis has more power (Peres-Neto and Jackson 2001), but its suitability is yet to be established on huge data sets. As a new addition to the debate about Mantel and Procrustes tests, Legendre, Fortin and Borcard (2015) recently emphasized the importance of linearity and homoscedascity assumptions of the Mantel test.

Commonly used ordination methods are often unable to separate effects of dispersion and locality, because they make wrong assumptions about how these vary in relation to each other (Warton, Wright and Wang 2012). This can be solved by model-based ordinations (Hui *et al.* 2015; Warton *et al.* 2015a) that use mixture models or latent variable models. Both of these are based on the GLM framework, and thus make explicit assumptions about the sampling distributions of the data. Latent variables summarize variation in species abundance and site properties, and can be used to produce ordinations. Model-based ordinations explicitly account for the statistical distribution of data and the mean–variance relationship (Fig. 3) thus, benefiting from the whole range of model diagnostic and model comparison tools. It is also possible to add predictors to models, allowing constraints similarly to CA (Warton *et al.* 2015a). Model-based ordinations regularly outperform algorithmic ordinations both in simulations, and biological data analysis (Hui *et al.* 2015; Warton *et al.* 2015a).

Changes in community composition

Multivariate hypothesis testing is required to evaluate how community composition changes with experimental treatments or environmental variation. It is challenging to simultaneously deal with all species, so either the dimensionality is first reduced and then hypotheses are tested (Legendre and Legendre 2012), or regressions are run for each species separately and then the dimensionality of the results is reduced by partitioning the community-level variation, while also considering the non-independence of species observations (Wang *et al.* 2012; Warton, Wright and Wang 2012).

The most widely used method to analyze community data by reducing the dimensionality first is permutational multivariate analysis of variance (PERMANOVA; Anderson 2001; McArdle and Anderson 2001). A matrix of distances between samples is first calculated from OTU observations and variation in this matrix is partitioned, according to several predictors and their interactions. PERMANOVA can be applied to hierarchical designs and random effects by constraining permutations within treatments/plots. Model selection tools can be used to evaluate the most informative predictors. Implementations can be found in the Permanova+ add-on (Anderson 2005) of the Primer software (Clarke and Gorley 2006) and in the adonis function of the vegan package (Oksanen *et al.* 2015) of R (R Core Team 2015). There are several issues with PERMANOVA: (i) loss of information due to the compaction of many OTU observations into either a single distance matrix, or into canonical functions; (ii) distance matrices cannot deal with overdispersion, which is typical for HTS and other ecological count data, (iii) assumptions about the relationship between mean and variance.

An alternative approach is canonical correspondence analysis (CCA), which is popular in plant ecology. In CCA, the correlations of ordination axes with explanatory variables are tested in an iterative process, using permutation tests. The software

CANOCO5 (Šmilauer and Lepš 2014) and the vegan package in R offer a variety of options for interaction terms, covariates and hierarchical analyses.

Fitting separate regressions first allows for hierarchical designs, interaction terms, model evaluation (residual and Q–Q plots) and model selection tools. Multiple regressions do not require data normalization (O'Hara and Kotze 2010; Warton, Wright and Wang 2012), because a link function defines the transformation of the mean into a linear function of the predictors, turning regressions into generalized linear models (GLMs). GLMs may account for the mean-variance relationship with a dispersion parameter according to the data. The mvabund package (Wang *et al.* 2012) simultaneously fits multispecies GLMs, partitions the effects of the predictors and computes test statistics that additionally account for correlations among OTUs.

Selection of predictors of community composition

When a large number of variables could affect community composition, a smaller number often must be selected. Ecological reasoning should always be used to reduce the number of predictors (e.g. using literature, theory, or preliminary data). After this, a model selection can be done in several ways. The classical approach is to add and remove predictors and compare models using significance tests or (more commonly) information criterion such as AIC. Other approaches use cross-validation (i.e. fitting models to part of the data and then testing the fit on the rest of the data). An alternative approach, such as lasso (Tibshirani 1994), constrains models to shrink effects towards zero, so variables with little effect are reduced to zero. However, many variable combinations may be roughly equally plausible (Whittingham *et al.* 2006), and therefore model selection is certainly not a substitute for ecological reasoning. Model selection works equally well for both continuous and discrete variables, and their mixtures.

Model selection is further complicated by multicollinearity among continuous predictors (Dormann *et al.* 2013). Thresholds of correlation coefficients can be used to exclude strongly correlated variables (commonly at $|r| > 0.7$), which is supported by simulations (Dormann *et al.* 2013). Correlation thresholds can be combined with stepwise model selection by discarding all predictors that are collinear with the best-fitting variables. An alternative approach is to summarize variables as principal coordinates (or components; PCs) using PCoA or PCA. The first few PCs are then used as predictors in models. This approach is mostly used for continuous data, although summarizing discrete data with PCA or PCoA will likely work. A mixture of continuous and discrete may be problematic. However, there is no reason why changes in community composition should be explained by variation summarized on the selected components (Hadi and Ling 1998). Further, it is not straightforward to understand what causes change in community composition as each component often summarizes variation from several predictors.

Most affected OTUs

Beyond the general changes in community composition, it may be important to understand which OTUs are specifically affected by predictors. OTU-level reactions may be estimated and visualized in the ordination space (e.g. in biplots). In CA, both OTU and sample scores can be calculated, so the OTU scores show which of these drive the divergence of samples (Šmilauer and Lepš 2014; Clemmensen *et al.* 2015). As described above, the mvabund package explicitly calculates multiple regressions and effect size of predictors for each OTU, so these can be compared directly.

Random Forest machine learning algorithms can be also used to model the environmental effects on the distribution of OTUs (Breiman 2001), being able to capture non-linear relationships without normality and distribution assumptions. McMurdie and Holmes (2014) discuss several tests inspired by RNA-sequencing data analysis for detecting OTUs affected by experimental factors. The recently developed ANCOM (analysis of composition of microbiomes; Mandal *et al.* 2015) compares the composition of microbiomes without explicit assumption about distributions with regard to environmental factors (but log-transforming the abundance data first). An alternative approach is to calculate statistics that measure the association between predictors and OTU incidences, either using correlations or indicator values (INDVAL; e.g. De Cáceres *et al.* 2012).

The relative effects of a predictor on OTUs can be used to rank the OTUs by their sensitivity (although such rankings are often not stable; see Goldstein and Spiegelhalter 1996). They can also be used to identify indicator OTUs. A comprehensive discussion of indicator species can be found in (Noss 1990; Rolstad *et al.* 2002; Favreau *et al.* 2006).

INFERENCE OF SPECIES INTERACTIONS FROM CO-OCCURRENCE DATA

Beyond the effects of the environment, the structure of communities is also influenced by interactions among species. Hypotheses about interactions may be derived from co-occurrence networks of species association. Because co-occurrence data (presence/absence or abundance) are inherently multivariate, the interactions are typically described by a number of indices that reduce the pattern down to a smaller number of statistics. In particular, modularity (i.e. where the associates can be grouped into discrete groups, with a few interactions between groups) and nestedness (where the composition of sites with fewer species tend to be subsets of species pools from more species-rich sites) are ecologically informative (Weitz *et al.* 2013). If a network is known (or has been inferred), then network properties—such as modularity, clustering coefficient, degree distribution or the average path length—can be calculated by several programs, e.g. the igraph package of R (Csárdi and Nepusz 2006) or Cytoscape NetworkAnalyzer (Shannon *et al.* 2003). The bipartite package of R (Dormann, Gruber and Fruend 2008; Dormann *et al.* 2009) and the BiMAT Matlab program (Flores *et al.* 2016) are tailored to the analysis of bipartite networks, such as host-parasite and host-symbiont associations.

A major problem is to infer the network from observational data. Some form of replication of observations of communities is needed to evaluate co-occurrences, either from a single sample from different sites or from repeated sampling at the same site. Hekstra and Leibler (2012) show that in the ideal case, the same species interactions can be inferred from spatial as well as temporal replication of the ecosystem under study. Co-occurrence patterns may be negative or positive, but any significant associations may stem from similar responses to the environment, spatial or temporal autocorrelation and/or interactions with another species. In the absence of experimental data, it is difficult to rule out hidden correlations among multiple species and unmeasured environmental variables.

Another important problem is that read counts are often rarefied or normalized to account for differences in sequencing depth (Friedman and Alm 2012). In this case, network inference

algorithms are dealing with proportional rather than absolute abundances. Correlations computed on proportional data can be severely distorted (Aitchison 2003). Microbial network inference tools address this problem by including compositionality-robust dissimilarity measures (e.g. CoNet, Faust and Raes 2012) or by analyzing relative abundances through a log-ratio transformation (e.g. SparCC—Friedman and Alm 2012; SPIEC-EASI—Kurtz *et al.* 2015). Model-based analyses can include a parameter that estimates the total abundance: when working with counts this is straightforward because of a simple relationship between the Poisson and multinomial distribution. See (Aitchison 2003) for more discussion about compositional data analysis.

Spatial replication

In case of spatial replication, co-variation in species occurrence or abundance can be examined with programs such as MENA (Deng *et al.* 2012), CoNet (Faust and Raes 2012), SparCC (Friedman and Alm 2012), LSA (Ruan *et al.* 2006; Xia et al., 2011, 2013; Durno *et al.* 2013) MIC (Reshef *et al.* 2011), REBACCA (Ban, An and Jiang 2015), CCLasso (Fang *et al.* 2015) and SPIEC-EASI (Kurtz *et al.* 2015). The performance of some of these tools was recently compared on synthetic data by Weiss *et al.* (2016), who suggests an ensemble approach to deal with the compositionality and sparsity issues of large OTU datasets. In all of these implementations, the potential effect of confounding factors remains neglected. However, if environmental factors have been measured, association of taxa resulting from an underlying environmental parameter can be identified with the interaction information (Lima-Mendez *et al.* 2015).

Recently, modeling tools have been developed to separate the responses of taxa to the environment from biotic interactions (Warton *et al.* 2015a). If we knew the environmental drivers, we could use single-species models to regress out environmental parameters. Correlations in the residual variation can then be used to infer interactions among taxa. In practice, the models used to do this simultaneously estimate the effects of the environment and the residual interactions. In the same way that GLMs are extensions of linear models, this is an extension of MANOVA models to non-normal responses. Non-linear dependencies can be also modeled when species do not respond linearly to the environment, i.e. organisms are often less abundant when environmental parameters such as pH or temperature deviate from their optima. Model-based interaction inferences can explicitly address both compositionality and sparsity, but they are yet to be compared to other approaches.

With many species, the full interaction matrix is large, with n(n-1)/2 parameters for n species. For example, for 10 000 species there are 49 995 000 terms in the matrix. This matrix has to be constrained in some way to make the estimation computationally feasible. One approach is to set some correlations to zero, but this can be difficult with so many terms. An alternative approach is to force the matrix to lie in a sub-space (i.e. to make it depend on a small number of factors). Hui *et al.* (2015) have developed an approach based on factor analysis, which reduces the matrix to a small number of factors. In essence, one can imagine that there are several unidentified environmental covariates that jointly affect the species, and the model estimates these 'latent factors' together with their effects on the species. So, if a latent factor affects two species in the same way, their presence will be positively correlated. In this way, the total number of parameters can be reduced, and if two factors are used, they can easily be visualized in a scatterplot.

Temporal replication

When replicate samples are taken from the same community at multiple times, we can investigate the temporal dynamics of the community. Methods for estimating correlations between pairs of time series have been developed, e.g. local similarity analysis (LSA), which employs a dynamic programming algorithm to identify lags between two time series (Ruan *et al.* 2006; Durno *et al.* 2013; Xia *et al.* 2013). LSA is widely applied to marine and freshwater time series data (Steele *et al.* 2011; Eiler, Heinrich and Bertilsson 2012; Gilbert *et al.* 2012; Sugihara *et al.* 2012: 2; Chow *et al.* 2014). Time series also allow the application of cross-prediction techniques (Granger 1969; Sugihara *et al.* 2012) and the detection of relationships in chaotic data (Sugihara *et al.* 2012). Furthermore, when time series are sufficiently long, the change of interaction patterns over time can be studied with time-varying networks (reviewed by Faust *et al.* 2015). The problems of temporally replicated samples including time series mirror those for spatial replicates, but the lagged time effects may be much stronger.

With time series data, community models such as the generalized Lotka–Volterra (gLV) model can be fit under the (strong) assumption that the dynamics are deterministic (Mounier *et al.* 2008; Marino *et al.* 2013; Stein *et al.* 2013; Fisher and Mehta 2014). Species abundances are described in terms of growth rates and interaction matrices. If we assume that the dynamics is stochastic, a variant of the neutral model (Sloan *et al.* 2006) or techniques from econometrics can be adapted. If we assume that the dynamics is linear (possibly on a transformed scale), we can use vector autoregressive models (Lütkepohl 2006). The first-order lags have been used to study co-occurrence (e.g. Mutshinda, O'Hara and Woiwod 2009). With large communities, the same problem of a large number of parameters arises. Time-lagged responses can be handled by adapting recent approaches for variable selection in regression, e.g. shrinkage or (in a Bayesian context) slab and spike priors (Mutshinda, O'Hara and Woiwod 2009) (Fig. 4). The residual covariance matrix is more difficult to simplify, although the latent variable approach of Hui *et al.* (2015) can be adapted, i.e. the residual covariance can be modelled in terms of a smaller number of variables. An alternative approach is dynamic factor analysis (Zuur *et al.* 2003), which models a smaller number of latent variables, which are then mapped onto the whole process. It is straightforward to add explanatory variables to these models.

STRUCTURAL EQUATION MODELING

The methods outlined above assume a simple relationship between the environment and microbes. However, the relationship may be more complex, with different aspects of the environment affecting other abiotic factors, as well as the biotic component. Thus, we may want to disentangle direct and indirect effects. This can be done using structural equation modeling (SEM; Bullock, Harlow and Mulaik 1994).

Structural equation models are increasingly used in the community ecology of macroorganisms to test hypotheses about causal relationships between environmental predictors and response variables (Scherber *et al.* 2010). To date only a few studies exist that apply them in microbial ecology to test causal hypotheses about the relationship between environmental predictors and response variables (Antoninka *et al.* 2009; Tedersoo *et al.* 2014; Xiang *et al.* 2014). The initial models are generated based on a priori assumptions about direct and indirect relationships among endogenous and exogenous variables. The best model

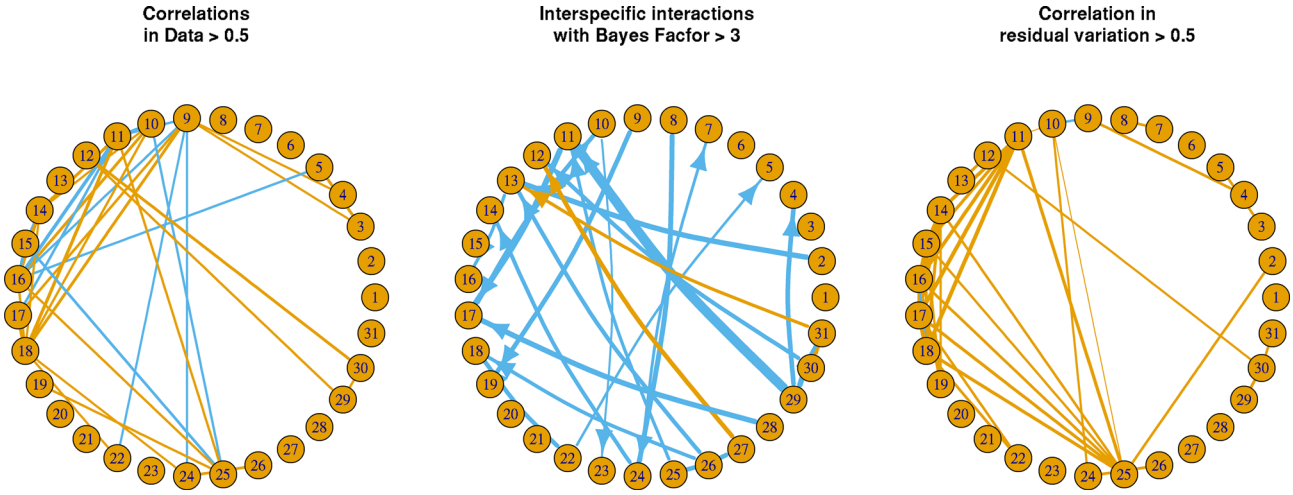

Figure 4. Time-lagged interactions inferred from co-occurrence time series. (**A**) Relatively strong ($|r| > 0.5$) correlations in the raw co-occurrence data. (**B**) The first-order time-lagged interactions between OTU read abundances. Interactions are not correlated with the raw correlations in the data. Only interactions with a Bayes factor indicative of positive evidence ($BF > 3$) are shown. All interactions are one way and arrows denote directional effects. (**C**) Correlations in the residual variation are correlated with correlations in the raw data ($r = 0.83$). Blue and brown lines indicate positive and negative links (associations), respectively. Link width is proportional with the correlation coefficient or Bayes factor. The interactions were estimated with frequent OTUs of the Western English Channel microbial community data set (Gilbert *et al.* 2012).

can be chosen by comparing the fit of different models which assume different causal relationships with several statistics, e.g. chi square, Akaike information criterion (AIC), comparative fit index (CFI) and the root mean square error of approximation (RMSEA). These statistics vary in their robustness, depending on sample size and model parameters (Hu and Bentler 1999). The optimal model can be found by stepwise addition or deletion of paths. If there are multiple dependent variables, individual models for each dependent variable can be optimized and integrated into a final model, following further model fit evaluation. With large models, the amount of data can become a problem: Grace *et al.* (2012) recommended at least 10 samples for each parameter in the model, with a minimum of 50 samples in total. SEMs are highly sensitive to excluded parameters (confounding variables). SEM analysis can be performed with AMOS and Mplus software (Byrne 2012), and OpenMx (Boker *et al.* 2011) in R.

FUTURE PERSPECTIVES

Methodological developments in bioinformatics and the statistical analysis of HTS-based marker gene data are being strongly influenced by a generation of new technologies. The earliest HTS papers reported a vast microbial diversity (Sogin *et al.* 2006), and we expect that fundamental biodiversity studies will remain a major research direction for some time. We expect improved ability to separate noise from signal in the rare biosphere (Delmont *et al.* 2011). This will build on current HTS data and models that can detect unlikely indels and substitutions in conserved primary structure and unfeasible changes to secondary structure of DNA/RNA. There will certainly be a growing appreciation that commonly used approaches for OTU delineation and classification (taxonomy assignment) blur or fail to capture a significant proportion of microbial diversity. Classification will certainly become more uniform, with phylotypes receiving unique identifiers (e.g. digital object identifiers; Tedersoo *et al.* 2015b), to allow for rapid cross-comparability among studies. These serve as a firm basis for metaanalyses of diversity and composition of microorganisms (Meiser, Bálint and Schmitt 2014). Functional diversity and its drivers will be increasingly emphasized as sequencing capacity and bioinformatics tools for metagenomics data further improve and capture probes for sets of multiple functional genes are developed (Manoharan *et al.* 2015). The continuous demand for improved metadata in marker gene databases (Yilmaz *et al.* 2011) and other developments in the functional annotation of OTUs is improving the way functional diversity is inferred from marker gene studies (Nguyen *et al.* 2016a), at least for eukaryotes, where horizontal gene transfer is relatively rare. This is further facilitated by *in situ* sequencing and annotating genomes, especially from common but unculturable groups (Sanli *et al.* 2015). Integrating 'omics approaches with marker gene-based community characterization will certainly lead to a better understanding of relationships between diversity, composition and function (Lindahl and Kuske 2013; Miki, Yokokawa and Matsui 2014; Peršoh 2015). Community phylogeny and comparative phylogenetics methods will help us improve our understanding of evolutionary ecology and the assembly of microbial communities. For fungi, long markers including variable and conserved parts, or methods for automated mapping of ITS-based OTUs to phylogenies are required for such approaches. In network analysis, computationally efficient methods accounting for other aspects such as phylogenetic relatedness, environmental variation and multiple interactions are urgently required. In metagenomic analyses, network approaches could link the observations of dominant taxa to specific functional genes or attributes in the community. Most importantly, microbial ecologists should become more aware of recent statistical developments by computer scientists, statisticians and ecologists and actively seek to implement or modify these for microbial community data. We hope that high-throughput marker gene studies will benefit from the statistical practices and perspectives evaluated here, and that the review will lead to a greater integration of microbial and community ecology.

ACKNOWLEDGEMENTS

The authors thank Imke Schmitt for constructive comments about the text and the discussed analyses.

FUNDING

MB and MU are supported by the German Research Foundation (DFG, grant BA 4843/2-1 to MB and UN 262/9-1 to MU). MÖ is supported by the Estonian Research Council (ERC; grants 9050, IUT20-28). LT receives funding from the ERC grants PUT171, 9286, EMP265 and EcolChange. MLS receives funding from the Alfred P. Sloan Foundation.

Conflict of interest. None declared.

REFERENCES

Abarenkov K, Tedersoo L, Nilsson RH *et al*. PlutoF–a web based workbench for ecological and taxonomic research, with an online implementation for fungal ITS sequences. *Evol Bioinform* 2010;**6**:189–96 (online).

Ainsworth TD, Krause L, Bridge T *et al*. The coral core microbiome identifies rare bacterial taxa as ubiquitous endosymbionts. *ISME J* 2015;**9**:2261–74.

Aitchison J. *A Concise Guide to Compositional Data Analysis*. Girona, Italy, 2003, http://ima.udg.edu/activitats/codawork05/A_concise_guide_to_compositional_data_analysis.pdf.

Amend AS, Seifert KA, Bruns TD. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Mol Ecol* 2010;**19**:5555–65.

Anderson MJ. A new method for non-parametric multivariate analysis of variance. *Austral Ecol* 2001;**26**:32–46.

Anderson MJ. Permutational multivariate analysis of variance. *Dep Stat Univ Auckl Auckl* 2005.

Anderson MJ, Walsh DCI. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? *Ecol Monogr* 2013;**83**:557–74.

Antoninka A, Wolf JE, Bowker M *et al*. Linking above- and belowground responses to global change at community and ecosystem scales. *Glob Change Biol* 2009;**15**:914–29.

Ashelford KE, Chuzhanova NA, Fry JC *et al*. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl Environ Microb* 2006;**72**:5734–41.

Bálint M, Bartha L, O'Hara RB *et al*. Relocation, high-latitude warming and host genetic identity shape the foliar fungal microbiome of poplars. *Mol Ecol* 2015;**24**:235–48.

Bálint M, Schmidt P-A, Sharma R *et al*. An Illumina metabarcoding pipeline for fungi. *Ecol Evol* 2014;**4**:2642–53.

Ban Y, An L, Jiang H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics* 2015;**31**:3322–9.

Baselga A. Partitioning the turnover and nestedness components of beta diversity. *Global Ecol Biogeogr* 2010;**19**:134–43.

Baselga A, Orme CDL. betapart: an R package for the study of beta diversity. *Method Ecol Evol* 2012;**3**:808–12.

Bik HM, Fournier D, Sung W *et al*. Intra-genomic variation in the ribosomal repeats of nematodes. *PLoS One* 2013;**8**:e78230.

Bik HM, Thomas WK. Metagenomics will highlight and drive links to taxonomic data: reply to Murray. *Trends Ecol Evol* 2012;**27**:652–3.

Bohannan BJM, Hughes J. New approaches to analyzing microbial biodiversity data. *Curr Opin Microbiol* 2003;**6**:282–7.

Boker S, Neale M, Maes H *et al*. OpenMx: an open source extended structural equation modeling framework. *Psychometrika* 2011;**76**:306–17.

Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.

Brodin J, Mild M, Hedskog C *et al*. PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. *PloS One* 2013;**8**:e70388.

Brown SP, Veach AM, Rigdon-Huss AR *et al*. Scraping the bottom of the barrel: are rare high throughput sequences artifacts? *Fungal Ecol* 2015;**13**:221–5.

Bullock HE, Harlow LL, Mulaik SA. Causation issues in structural equation modeling research. *Struct Equ Modeling* 1994;**1**:253–67.

Burbank DW, Anderson RS. *Tectonic Geomorphology*. Chichester, West Sussex; Hoboken, NJ: J. Wiley & Sons, 2012.

Buttigieg PL, Ramette A. A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol Ecol* 2014;**90**:543–50.

Byrne BM. *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming*. Oxon, UK: Routledge, 2012.

Caporaso JG, Kuczynski J, Stombaugh J *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Meth* 2010;**7**:335–6.

Carlsen T, Aas AB, Lindner D *et al*. Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecol* 2012;**5**:747–9.

Chao A, Colwell RK, Lin C-W *et al*. Sufficient sampling for asymptotic minimum species richness estimators. *Ecology* 2009;**90**:1125–33.

Chao A, Jost L. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* 2012;**93**:2533–47.

Chase JM, Kraft NJB, Smith KG *et al*. Using null models to disentangle variation in community dissimilarity from variation in α-diversity. *Ecosphere* 2011;**2**:art24.

Chow C-ET, Kim DY, Sachdeva R *et al*. Top-down controls on bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists. *ISME J* 2014;**8**:816–29.

Clarke K, Gorley R. *PRIMER v7: User Manual/Tutorial*. Plymouth,UK: PRIMER-E, 2006.

Clemmensen KE, Finlay RD, Dahlberg A *et al*. Carbon sequestration is related to mycorrhizal fungal community shifts during long-term succession in boreal forests. *New Phytol* 2015;**205**:1525–36.

Cole JR, Wang Q, Cardenas E *et al*. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 2009;**37**:D141–5.

Colwell RK, Chao A, Gotelli NJ *et al*. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J Plant Ecol* 2012;**5**:3–21.

Colwell RK, Mao CX, Chang J. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology* 2004;**85**:2717–27.

Csárdi G, Nepusz T. The igraph software package for complex network research. *InterJournal* Complex Systems 2006;**1695**.

De Cáceres M, Legendre P, Wiser SK *et al*. Using species combinations in indicator value analyses. *Methods Ecol Evol* 2012;**3**:973–82.

Delmont TO, Malandain C, Prestat E *et al*. Metagenomic mining for microbiologists. *ISME J* 2011;**5**:1837–43.

Deng Y, Jiang YH, Yang Y *et al*. Molecular ecological network analyses. *BMC Bioinformatics* 2012;**13**:113+.

DeSantis TZ, Hugenholtz P, Larsen N *et al*. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microb* 2006;**72**:5069–72.

Doolittle WF, Zhaxybayeva O. On the origin of prokaryotic species. *Genome Res* 2009;**19**:744–56.

Dormann CF, Elith J, Bacher S *et al.* Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 2013;**36**:27–46.

Dormann CF, Frueund J, Bluethgen N *et al.* Indices, graphs and null models: analyzing bipartite ecological networks. *Open Ecol J* 2009;**2**:7–24.

Dormann CF, Gruber B, Fruend J. Introducing the bipartite package: analysing ecological networks. *R News* 2008;**8**: 8–11.

Durno WE, Hanson NW, Konwar KM *et al.* Expanding the boundaries of local similarity analysis. *BMC Genomics* 2013; **14**:S3.

Eickbush TH, Eickbush DG. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* 2007;**175**:477–85.

Eiler A, Heinrich F, Bertilsson S. Coherent dynamics and association networks among lake bacterioplankton taxa. *ISME J* 2012;**6**:330–42.

Epstein S. The phenomenon of microbial uncultivability. *Curr Opin Microbiol* 2013;**16**:636–42.

Eren AM, Maignien L, Sul WJ *et al.* Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol* 2013;**4**:1111–9.

Eren AM, Morrison HG, Lescault PJ *et al.* Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* 2015a;**9**:968–79.

Eren AM, Sogin ML, Morrison HG *et al.* A single genus in the gut microbiome reflects host preference and specificity. *ISME J* 2015b;**9**:90–100.

Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. *Science* 2008;**320**:1034–9.

Fang H, Huang C, Zhao H *et al.* CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* 2015;**31**:3172–80.

Faust K, Lahti L, Gonze D *et al.* Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr Opin Microbiol* 2015;**25**:56–66.

Faust K, Raes J. Microbial interactions: from networks to models. *Nat Rev Microbiol* 2012;**10**:538–50.

Favreau JM, Drew CA, Hess GR *et al.* Recommendations for assessing the effectiveness of surrogate species approaches. *Biodivers Conserv* 2006;**15**:3949–69.

Ficetola G, Coissac E, Zundel S *et al.* An In silico approach for the evaluation of DNA barcodes. *BMC Genomics* 2010;**11**:434+.

Fierer N, Bradford MA, Jackson RB. Toward an ecological classification of soil bacteria. *Ecology* 2007;**88**:1354–64.

Fisher CK, Mehta P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One* 2014;**9**:e102451.

Flores CO, Poisot T, Valverde S *et al.* BiMat: a MATLAB package to facilitate the analysis of bipartite networks. *Methods Ecol Evol* 2016;**7**:127–32.

Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 2012;**8**:e1002687.

Fujita MK, Leaché AD, Burbrink FT *et al.* Coalescent-based species delimitation in an integrative taxonomy. *Trends Ecol Evol* 2012;**27**:480–8.

Gevers D, Cohan FM, Lawrence JG *et al.* Opinion: re-evaluating prokaryotic species. *Nat Rev Microbiol* 2005;**3**:733–9.

Gilbert JA, Steele JA, Caporaso JG *et al.* Defining seasonal marine microbial community dynamics. *ISME J* 2012;**6**:298–308.

Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Stat Soc Ser A Stat Soc* 1996:385–443.

Gotelli NJ, Colwell RK. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol Lett* 2001;**4**:379–91.

Grace JB, Keeley JE, Johnson DJ *et al.* Structural equation modeling and the analysis of long-term monitoring data. In: Gitzen RA, Millspaugh JJ, Cooper AB *et al.* (eds). *Design and Analysis of Long-Term Ecological Monitoring Studies*. Cambridge, England: Cambridge University Press, 2012, 325–58.

Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 1969;**37**:424–38.

Guillot G, Rousset F. Dismantling the Mantel tests. *Methods Ecol Evol* 2013;**4**:336–44.

Gwinn DC, Allen MS, Bonvechio KI *et al.* Evaluating estimators of species richness: the importance of considering statistical error rates. *Methods Ecol Evol* 2016;**7**:294–302.

Hadi AS, Ling RF. Some cautionary notes on the use of principal components regression. *Am Stat* 1998;**52**:15–9.

Haegeman B, Hamelin J, Moriarty J *et al.* Robust estimation of microbial diversity in theory and in practice. *ISME J* 2013;**7**:1092–101.

Hamady M, Lozupone C, Knight R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* 2009;**4**:17–27.

Hao X, Jiang R, Chen T. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* 2011;**27**:611–8.

He Y, Caporaso JG, Jiang X-T *et al.* Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* 2015;**3**:20.

Hekstra DR, Leibler S. Contingency and statistical laws in replicate microbial closed ecosystems. *Cell* 2012;**149**:1164–73.

Hill MO. Diversity and evenness: a unifying notation and its consequences. *Ecology* 1973;**54**:427–32.

Hinchliff CE, Smith SA, Allman JF *et al.* Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *P Natl Acad Sci USA* 2015;**112**:12764–9.

Hong S-H, Bunge J, Jeon S-O *et al.* Predicting microbial species richness. *P Natl Acad Sci USA* 2006;**103**:117–22.

Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Modeling* 1999;**6**:1–55.

Hughes JB, Hellmann JJ, Ricketts TH *et al.* Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microb* 2001;**67**:4399–406.

Hui FKC, Taskinen S, Pledger S *et al.* Model-based approaches to unconstrained ordination. *Methods Ecol Evol* 2015;**6**: 399–411.

Huse SM, Welch DBM, Voorhis A *et al.* VAMPS: a website for visualization and analysis of microbial population structures. *BMC Bioinformatics* 2014;**15**:1–7.

Huse SM, Welch DM, Morrison HG *et al.* Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* 2010;**12**:1889–98.

Huson DH, Mitra S, Ruscheweyh H-J *et al.* Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 2011;**21**:1552–60.

Ihrmark K, Bödeker ITM, Cruz-Martinez K *et al.* New primers to amplify the fungal ITS2 region – evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiol Ecol* 2012;**82**:666–77.

Jansson JK, Prosser JI. Microbiology: the life beneath our feet. *Nature* 2013;**494**:40–1.

Koeppel AF, Wu M. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Res* 2013;**41**:5175–88.

Kõljalg U, Nilsson RH, Abarenkov K *et al.* Towards a unified paradigm for sequence-based identification of fungi. *Mol Ecol* 2013;**22**:5271–7.

Kuczynski J, Liu Z, Lozupone C *et al.* Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods* 2010;**7**:813–9.

Kunin V, Engelbrektson A, Ochman H *et al.* Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* 2010;**12**:118–23.

Kurtz ZD, Müller CL, Miraldi ER *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* 2015;**11**:e1004226.

Legendre P, Fortin M-J, Borcard D. Should the Mantel test be used in spatial analysis? *Methods Ecol Evol* 2015;**6**:1239–47.

Legendre P, Legendre L. *Numerical Ecology*, **Vol. 24**, 3rd edn. Amsterdam: Elsevier, 2012.

Leininger S, Urich T, Schloter M *et al.* Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* 2006;**442**:806–9.

Lima-Mendez G, Faust K, Henry N *et al.* Determinants of community structure in the global plankton interactome. *Science* 2015;**348**:1262073.

Lindahl BD, Kuske CR. Metagenomics for study of fungal ecology. In: Francis M (ed). *The Ecological Genomics of Fungi*. Hoboken, NJ: John Wiley & Sons, Inc, 2013, 279–303.

Locey KJ, Lennon JT. *Scaling laws predict global microbial diversity*. 2016, DOI: 10.7287/peerj.preprints.1451v3.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.

Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microb* 2005;**71**:8228–35.

Lozupone CA, Hamady M, Kelley ST *et al.* Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microb* 2007;**73**:1576–85.

Lütkepohl H. Structural vector autoregressive analysis for cointegrated variables. In: Hübler PDO, Frohn PDJ (eds). *Modern Econometric Analysis*. Berlin Heidelberg: Springer, 2006, 73–86.

McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 2001;**82**:290–7.

McGill BJ, Etienne RS, Gray JS *et al.* Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol Lett* 2007;**10**:995–1015.

McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 2014;**10**:e1003531.

Magurran AE. *Measuring Biological Diversity*. Malden, MA: John Wiley & Sons, 2003.

Magurran AE. How ecosystems change. *Science* 2016;**351**:448–9.

Magurran AE, Dornelas M, Moyes F *et al.* Rapid biotic homogenization of marine fish assemblages. *Nat Commun* 2015;**6**:8405.

Mahé F, Rognes T, Quince C *et al.* Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2014;**2**:e593.

Mandal S, Van Treuren W, White RA *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* 2015;**26**:27663.

Manly BFJ, Alberto JAN. *Introduction to Ecological Sampling*. Boca Raton, FL: CRC Press, 2014.

Manoharan L, Kushwaha SK, Hedlund K *et al.* Captured metagenomics: large-scale targeting of genes based on "sequence capture" reveals functional diversity in soils. *DNA Res* 2015;**22**:451–60.

Margesin R, Miteva V. Diversity and ecology of psychrophilic microorganisms. *Res Microbiol* 2011;**162**:346–61.

Marino S, Baxter NT, Huffnagle GB *et al.* Mathematical modeling of primary succession of murine intestinal microbiota. *P Natl Acad Sci USA* 2013;**111**:439–44.

Matsen FA, Evans SN. Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PLoS One* 2013;**8**, DOI: 10.1371/journal.pone.0056859.

Meiser A, Bálint M, Schmitt I. Meta-analysis of deep-sequenced fungal communities indicates limited taxon sharing between studies and the presence of biogeographic patterns. *New Phytol* 2014;**201**:623–35.

Miki T, Yokokawa T, Matsui K. Biodiversity and multifunctionality in a microbial community: a novel theoretical approach to quantify functional redundancy. *P Roy Soc Lond B Biol* 2014;**281**:20132498.

Mounier J, Monnet C, Vallaeys T *et al.* Microbial interactions within a cheese microbial community. *Appl Environ Microb* 2008;**74**:172–81.

Mutshinda CM, O'Hara RB, Woiwod IP. What drives community dynamics? *P Roy Soc Lond B Biol* 2009;**276**:2923–9.

Nguyen NH, Song Z, Bates ST *et al.* FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecol* 2016a;**20**:241–8.

Nguyen N-P, Warnow T, Pop M *et al.* A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *Npj Biofilms Microbiomes* 2016b;**2**:16004.

Noss RF. Indicators for monitoring biodiversity: a hierarchical approach. *Conserv Biol* 1990;**4**:355–64.

O'Hara RB. Species richness estimators: how many species can dance on the head of a pin? *J Anim Ecol* 2005;**74**:375–86.

O'Hara RB, Kotze DJ. Do not log-transform count data. *Methods Ecol Evol* 2010;**1**:118–22.

Oksanen J, Blanchet FG, Kindt R *et al. Vegan: Community Ecology Package*. 2015, http://CRAN.R-project.org/package=vegan.

Oliver AK, Brown SP, Callaham MA, Jr *et al.* Polymerase matters: non-proofreading enzymes inflate fungal community richness estimates by up to 15%. *Fungal Ecol* 2015;**15**:86–9.

Öpik M, Davison J, Moora M *et al.* DNA-based detection and identification of Glomeromycota: the virtual taxonomy of environmental sequences. *Botany* 2014;**92**:135–47.

Pace NR. A molecular view of microbial diversity and the biosphere. *Science* 1997;**276**:734–40.

Parada A, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time-series and global field samples. *Environ Microbiol* 2015, DOI: 10.1111/1462-2920.13023.

Patin NV, Kunin V, Lidström U *et al.* Effects of OTU clustering and PCR artifacts on microbial diversity estimates. *Microb Ecol* 2013;**65**:709–19.

Peres-Neto PR, Jackson DA. How well do multivariate data sets match? the advantages of a procrustean superimposition approach over the mantel test. *Oecologia* 2001;**129**:169–78.

Perisin M, Vetter M, Gilbert JA *et al.* 16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies. *ISME J* 2016;**10**:1020–4.

Peršoh D. Plant-associated fungal communities in the light of meta'omics. *Fungal Divers* 2015;**75**:1–25.

Preheim SP, Perrotta AR, Martin-Platero AM *et al.* Distribution-based clustering: using ecology to refine the operational taxonomic unit. *Appl Environ Microb* 2013;**79**:6593–603.

Prosser JI, Bohannan BJM, Curtis TP *et al.* The role of ecological theory in microbial ecology. *Nat Rev Microbiol* 2007;**5**: 384–92.

Pruesse E, Quast C, Knittel K *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007;**35**:7188–96.

Quince C, Curtis TP, Sloan WT. The rational exploration of microbial diversity. *ISME J* 2008;**2**:997–1006.

Quince C, Lanzen A, Curtis TP *et al.* Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 2009;**6**:639–41.

Quinn GP, Keough MJ. *Experimental Design and Data Analysis for Biologists*. Country Cambridge, UK: Cambridge University Press, 2002.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2015.

Reshef DN, Reshef YA, Finucane HK *et al.* Detecting novel associations in large data sets. *Science* 2011;**334**:1518–24.

Risso D, Ngai J, Speed TP *et al.* Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 2014;**32**:896–902.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.

Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010;**11**:R25.

Rolstad J, Gjerde I, Gundersen VS *et al.* Use of indicator species to assess forest continuity: a critique. *Conserv Biol* 2002;**16**: 253–7.

Rosling A, Cox F, Cruz-Martinez K *et al.* Archaeorhizomycetes: unearthing an ancient class of ubiquitous soil fungi. *Science* 2011;**333**:876–9.

Ruan Q, Dutta D, Schwalbach MS *et al.* Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* 2006;**22**:2532–8.

Ryberg M. Molecular operational taxonomic units as approximations of species in the light of evolutionary models and empirical data from Fungi. *Mol Ecol* 2015;**24**:5770–7.

Sanli K, Bengtsson-Palme J, Nilsson RH *et al.* Metagenomic sequencing of marine periphyton: taxonomic and functional insights into biofilm communities. *Aquat Microbiol* 2015:**6**:1192.

Scherber C, Eisenhauer N, Weisser WW *et al.* Bottom-up effects of plant diversity on multitrophic interactions in a biodiversity experiment. *Nature* 2010;**468**:553–6.

Schloss PD, Westcott SL, Ryabin T *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microb* 2009;**75**:7537–41.

Schmidt TSB, Matias Rodrigues JF, von Mering C. Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ Microbiol* 2015;**17**: 1689–706.

Schnell IB, Bohmann K, Gilbert MTP. Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol Ecol Resour* 2015;**15**:1289–303.

Shannon P, Markiel A, Ozier O *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.

Sipos R, Székely AJ, Palatinszky M *et al.* Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol* 2007;**60**:341–50.

Sloan WT, Lunn M, Woodcock S *et al.* Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ Microbiol* 2006;**8**:732–40.

Smets W, Leff JW, Bradford MA *et al.* A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. 2015:e1622.

Šmilauer P, Lepš J. *Multivariate Analysis of Ecological Data Using Canoco* 5, 2nd edn. Cambridge, UK: Cambridge University Press, 2014.

Sogin ML, Morrison HG, Huber JA *et al.* Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *P Natl Acad Sci USA* 2006;**103**:12115–20.

Steele JA, Countway PD, Xia L *et al.* Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J* 2011;**5**:1414–25.

Stegen JC, Lin X, Fredrickson JK *et al.* Quantifying community assembly processes and identifying features that impose them. *ISME J* 2013;**7**:2069–79.

Stein RR, Bucci V, Toussaint NC *et al.* Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol* 2013;**9**:e1003388.

Stingl U, Cho J-C, Foo W *et al.* Dilution-to-extinction culturing of psychrotolerant planktonic bacteria from permanently ice-covered lakes in the McMurdo Dry Valleys, Antarctica. *Microb Ecol* 2008;**55**:395–405.

Sugihara G, May R, Ye H *et al.* Detecting causality in complex ecosystems. *Science* 2012;**338**:496–500.

Taylor LR. Aggregation, variance and the mean. *Nature* 1961;**189**:732–5.

Tedersoo L, Anslan S, Bahram M *et al.* Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys* 2015;**10**:1–43.

Tedersoo L, Bahram M, Põlme S *et al.* Global diversity and geography of soil fungi. *Science* 2014;**346**:1256688.

Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B* 1994;**58**:267–88.

Tikhonov M, Leach RW, Wingreen NS. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J* 2015;**9**:68–80.

Unterseher M, Schnittler M, Dormann C *et al.* Application of species richness estimators for the assessment of fungal diversity. *FEMS Microbiol Lett* 2008;**282**:205–13.

Vamosi SM, Heard SB, Vamosi JC *et al.* Emerging patterns in the comparative analysis of phylogenetic community structure. *Mol Ecol* 2009;**18**:572–92.

van der Heijden MGA, Bardgett RD, van Straalen NM. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecol Lett* 2008;**11**: 296–310.

van der Heijden MGA, Klironomos JN, Ursic M *et al.* Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* 1998;**396**:69–72.

Wang GC, Wang Y. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl Environ Microb* 1997;**63**:4645–50.

Wang Y, Naumann U, Wright ST *et al*. mvabund– an R package for model-based analysis of multivariate abundance data. *Methods Ecol Evol* 2012;**3**:471–4.

Warton DI, Blanchet FG, O'Hara RB *et al*. So many variables: joint modeling in community ecology. *Trends Ecol Evol* 2015a;**30**:766–79.

Warton DI, Foster SD, De'ath G *et al*. Model-based thinking for community ecology. *Plant Ecol* 2015b;**216**:669–82.

Warton DI, Wright ST, Wang Y. Distance-based multivariate analyses confound location and dispersion effects. *Methods Ecol Evol* 2012;**3**:89–101.

Weiss S, Van Treuren W, Lozupone C *et al*. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J* 2016, DOI: 10.1038/ismej.2015.235.

Weiss SJ, Xu Z, Amir A *et al*. Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. 2015, DOI: 10.1038/ismej.2015.235.

Weitz JS, Poisot T, Meyer JR *et al*. Phage–bacteria infection networks. *Trends Microbiol* 2013;**21**:82–91.

Westcott SL, Schloss PD. *De novo* clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 2015;**3**:e1487.

Whittingham MJ, Stephens PA, Bradbury RB *et al*. Why do we still use stepwise modelling in ecology and behaviour? *J Anim Ecol* 2006;**75**:1182–9.

Xia LC, Ai D, Cram J *et al*. Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics* 2013;**29**:230–7.

Xia LC, Steele JA, Cram JA *et al*. Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst Biol* 2011;**5**:S15.

Xiang D, Verbruggen E, Hu Y *et al*. Land use influences arbuscular mycorrhizal fungal communities in the farming–pastoral ecotone of northern China. *New Phytol* 2014;**204**:968–78.

Yilmaz P, Kottmann R, Field D *et al*. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* 2011;**29**:415–20.

Zhang J, Kapli P, Pavlidis P *et al*. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* 2013;**29**:2869–76.

Zuur AF, Fryer RJ, Jolliffe IT *et al*. Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics* 2003;**14**:665–85.